5th IEEE Workshop on SIGNAL PROPAGATION ON INTERCONNECTS, Venice, Italy May, 13-16, 2001

3-D Heterogeneous ICs: A Technology for the Next Decade and Beyond

Kaustav Banerjee, Shukri J. Souri, Pawan Kapur and Krishna C. Saraswat

Center for Integrated Systems, Stanford University, Stanford, CA 94305
kaustav@cis.stanford.edu

Performance and cost of deep submicron VLSI is being increasingly dominated by the interconnects due to decreasing wire pitch and increasing die size. Additionally, heterogeneous integration of different technologies in one single chip is becoming increasingly desirable, for which planar (2-D) ICs may not be suitable. This paper gives an overview of the performance analysis of 3-D ICs, an attractive alternative architecture that can alleviate the interconnect related problems such as delay and power dissipation, and can also facilitate integration of heterogeneous technologies such as SOI, SiGe HBTs, GaAs and so on, in one single chip.

Continuous scaling of VLSI circuits is reducing gate delays but rapidly increasing interconnect delays. Semiconductor Industry Association roadmap [1] predicts that, beyond the 130 nm technology node, performance improvement of advanced VLSI is likely to begin to saturate unless a paradigm shift from present IC architecture is introduced. **Figure 1** illustrates this problem, where the gate delay and the optimized interconnect delay are shown as functions of various technology nodes based on [1]. Also, increasing interconnect loading affects the power consumption in high-performance chips. In fact, a significant fraction of the total chip power consumption can be due to the wiring network used for clock distribution, which is usually realized using long global wires. Additionally, interconnect scaling has significant implications for traditional computer-aided-design methodologies and tools which are causing the design cycles to increase, thus increasing the time-to-market and the cost per chip function. Furthermore, increasing drive for the integration of disparate signals and technologies is introducing various System-on-a-Chip (SoC) design concepts, for which existing planar (2-D) IC design may not be suitable.

At the 250 nm technology node, Copper with low-k dielectric was introduced to alleviate the adverse effects of increasing interconnect delay. However, as shown in Figure 1, below 130 nm technology node, substantial interconnect delays will result in spite of introducing these new materials, which in turn will severely limit the chip performance. Further reduction in interconnect delay cannot be achieved by introducing any new materials. Furthermore, with the aggressive scaling suggested by the ITRS [1], new physical and technological effects start dominating interconnect properties (**Figure 2**), which can further increase interconnect delays [2]. Additionally, the use of repeaters to lower the interconnect delay becomes increasingly futile as these large repeaters take up a lot of active silicon area (**Figure 3**). Also, the vias that connect such repeaters to the top global interconnect layers cause blockage in all the lower metal layers, thereby taking up substantial routing area in all metal tiers.

3-D integration (schematically illustrated in **Figure 4**) to create multi active layer ICs is a concept that can significantly improve deep submicron interconnect performance, increase transistor packing density, and reduce chip area and power dissipation [2], [3]. In the 3-D design architecture an entire (2-D) chip can divided into a number of blocks, and each block is placed on a separate active layer (Si or other material) that are stacked on top of each other. Each active layer in the 3-D structure can have multiple levels of interconnect. Each of these active layers can be connected together with short vertical inter-layer interconnects (VILICs) as shown in Figure 4. These VILICs can eliminate the long global wires that realize the inter-block communications in 2-D ICs. The 3-D architecture offers extra flexibility in system design, placement and routing. For instance, logic gates on a critical path can be placed very close to each other using multiple active layers. This would result in a significant reduction in RC delay, and can greatly enhance the performance of logic circuits. Furthermore, the 3-D chip design technology can be exploited to build SoCs by placing circuits with different voltage and performance requirements in different layers (**Figure 5**). For example, the digital and analog components in the mixed-signal systems can be placed on different Si layers thereby achieving better noise performance due to lower electromagnetic interference between such circuit blocks. From an heterogeneous integration point of view, mixed-technology assimilation could be made less complex and more cost effective by fabricating such technologies on separate substrates followed by physical bonding. At present, the technologies being investigated to realize 3-D structures are [2]: seeded recrystallization of amorphous Si, processed wafer bonding and silicon epitaxial growth.

Figure 6 shows the wire length distributions for 2-D and 3-D ICs with two active layers using ITRS data for the high-performance 50 nm technology node. It can be observed that the wiring requirement is significantly reduced for the global wires in 3-D ICs. This is due to the fact that these long wires have been converted to short VILICs as schematically illustrated in Figure 4. **Figure 7** shows area and performance optimization for 2-D and 2-layer-3-D ICs. This also illustrates how the optimum normalized semi-global pitch (associated with the minimum chip area) can be increased to obtain higher operating frequencies. Beyond the maximum performance point (6 GHz) for the 3-D chip in Figure 7 (normalized semiglobal pitch $\cong$1.75), the performance gain becomes increasingly smaller in comparison to the performance degradation resulting from the increase in chip area (or interconnect delay), and the clock frequency saturates (**Figure 8**). Also, the decrease in interconnect delay becomes progressively smaller as the number of active layers increases (**Figure 9**). This is due to the fact that area required by the VILICs begins to offset any (footprint) area saving due to increasing the number of active layers. **Figure 10** summarizes the performance improvement of the 3-D ICs.

Finally, for 3-D ICs involving high performance circuits, thermal issues need careful considerations [4]. **Figure 11(a)** shows that for most practical 3-D ICs (with up to 5 layers), the temperature of the layers increase linearly with the layer number, due to the dominating influence of the package thermal resistance on the chip (layer) temperature. **Figure 11(b)** shows that advancement in packaging/cooling solutions will be needed for high-performance 3-D structures.

Acknowledgements

This work is being supported by the MARCO Interconnect Focus Center and DARPA.

References

[1] "The International Technology Roadmap for Semiconductors" 1999.

[2] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: A Novel Chip design for Improving Deep Submicron Interconnect Performance and Systems-on-Chip Integration," *to appear in Proc. of the IEEE*, May, 2001.

[3] S. J. Souri, K. Banerjee, A. Mehrotra and K. C. Saraswat, "Multiple Si Layer ICs: Motivation, Performance Analysis, and Design Implications," *DAC*, 2000, pp. 213-220.

[4] S. Im and K. Banerjee, "Full Chip Thermal Analysis of Planar (2-D) and Vertically Integrated (3-D) High Performance ICs," *IEDM*, 2000, pp. 727-730.

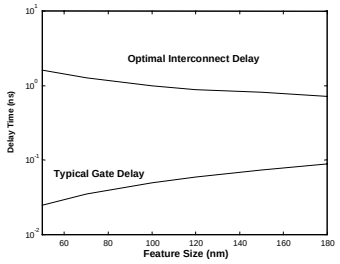


Figure 1. Typical gate and interconnect delays as functions of technology nodes. The interconnect delay assumes an optimally repeatered line and includes the delay due to the repeaters.

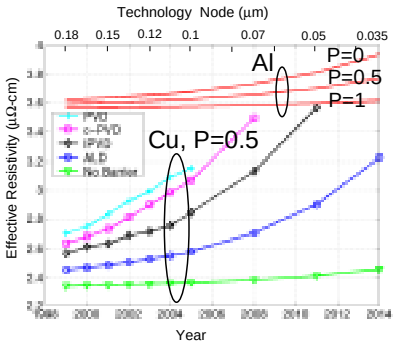


Figure 2. Effective resistivity of Cu lines (calculated with both scattering and barrier effects at 100 ^{0}C and for P = 0.5) as a function of technology node based on ITRS [1], for various barrier deposition technologies. Resistivity of Al interconnects are also shown for different values of the scattering parameter, P.

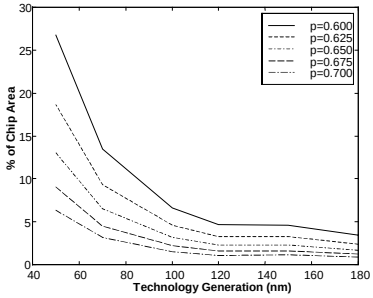


Figure 3. Percentage of chip area utilized by the repeaters as a function of various technology nodes, and Rent's exponent, p [2].

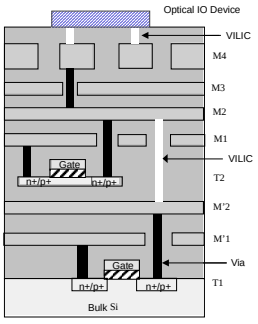


Figure 4. Schematic representation of 3-D integration with multilevel wiring network and VILICs. T1: first active layer device, T2: second active layer device, Optical I/O device: third active layer I/O device. M'1 and M'2 are for T1, M1 and M2 are for T2. M3 and M4 are shared by T1, T2, and the I/O device.

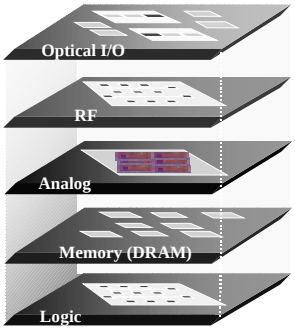


Figure 5. Schematic of a 3-D chip showing integrated heterogeneous technologies.

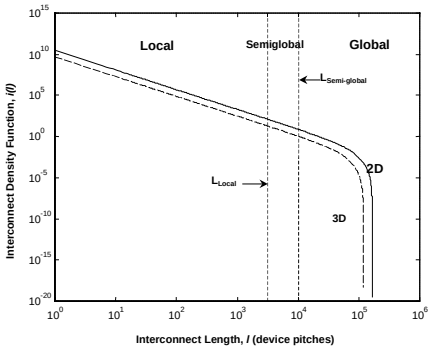


Figure 6. Wire-length distributions for the 2-D and a 2-layer 3-D IC. 3-D significantly reduces requirement for longest wires. Metal tiers determined by L_{Loc} and $L_{Semi\text{-}global}$ boundaries [2].

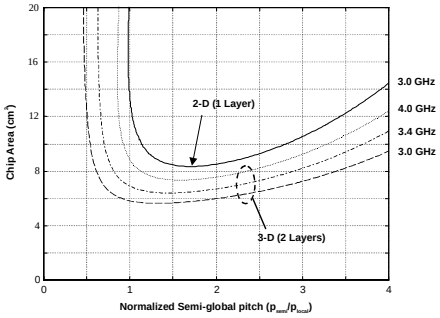


Figure 7. 3-D chip operating frequency (performance) increases with increases in semi-global wiring pitch. Chip (footprint) area also increases but remains below the 2-D chip area. If 3-D footprint area is made equal to 2-D chip area (= 8.17 cm^2 at the 50 nm node), an operating frequency of 6 GHz can be obtained for the 3-D chip.

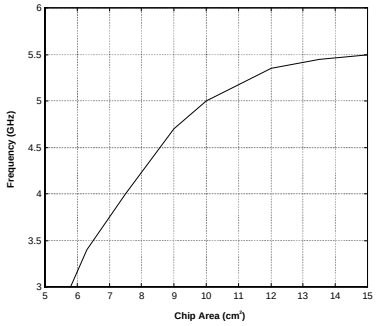


Figure 8. Performance improvement with increasing chip area for a two-layer 3-D IC. Total chip area is increased due to increasing wire pitch.

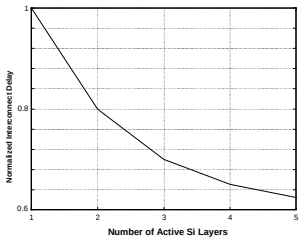


Figure 9. Interconnect delay normalized to single layer delay as a function of the number of active Si layers shown for 50 nm node. The VILICs are assumed to consume lateral area.

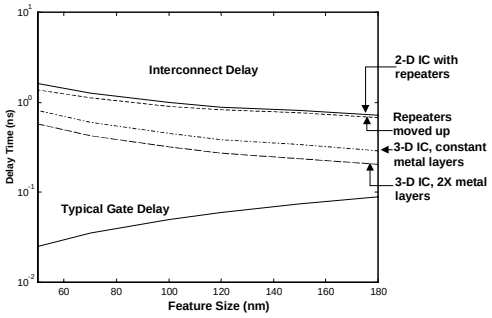
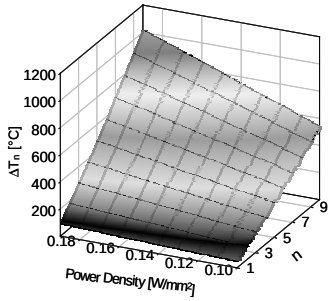


Figure 10. Comparison of interconnect delay as a function of technology nodes for 2-D and two-layer 3-D ICs. Moving repeaters to the upper active layer reduces interconnect delay by 9%. For the 50 nm node, 3-D (with same number of interconnect levels as the 2-D chip) shows significant delay reduction (63%). Increasing the number of metal levels in 3-D reduces interconnect delay by a further 35%. This figure is based on the assumption that 3-D chip (footprint) area equals 2-D chip area.



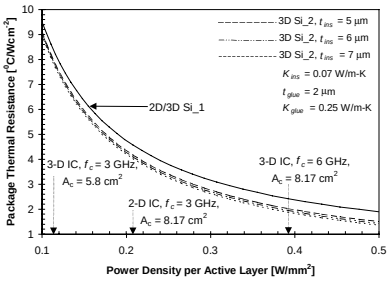


Figure 11. a) Temperature increase as a function of number of active layers (n) and the power density in each layer, b) required package thermal resistance for 2-D and two-layer 3-D ICs to maintain the temperature of any layer at 120 ^{0}C as a function of power density per layer. Heat sink is assumed at one end of the chip only. For the 3-D IC, as the dielectric thickness between the two active layers (t_{ins}) increases, lower values of the package thermal resistances are needed to maintain the temperature of the second active layer at 120 ^{0}C.